Assessing Large Language Models in Building a Structured Dataset from Reddit Data: A Methodological Study

Quinn Snell, Chase Westhoff, John Westhoff, Ethan Low, Carl Hanson, Shannon Tass

Submitted to: Journal of Medical Internet Research
on: March 17, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Assessing Large Language Models in Building a Structured Dataset from Reddit Data: A Methodological Study

Quinn Snell[1*] PhD; Chase Westhoff[1*] MS; John Westhoff[2*] MD, MPH; Ethan Low[1*]; Carl Hanson[1*] PhD; Shannon Tass[1*] PhD

[1]3361 TMCB Brigham Young University Provo US
[2] University of Nevada, Reno Reno US
[*]these authors contributed equally

Corresponding Author:
Quinn Snell PhD
3361 TMCB
Brigham Young University
3361 TMCB
Provo
US

Abstract

Background: In an era marked by the blooming reliance on digital platforms for healthcare consultation, the subreddit r/AskDocs has emerged as a pivotal forum. However, the vast, unstructured nature of forum data presents a formidable challenge; the extraction and meaningful analysis of such data require advanced tools that can navigate the complexities of language and context inherent in user-generated content.

Objective: Our objective was to evaluate employing Large Language Models (LLMs) to systematically transform the rich, unstructured textual data from AskDocs into a structured dataset, an approach that aligns more closely with human cognitive processes compared to traditional data extraction methods.

Methods: We developed a dataset of Reddit posts from r/AskDocs by extracting key information via human annotators. Then using specially engineered prompts we used state-of-the-art Large Language Models (LLMs) to extract data from posts and compared the results. The variation in the LLMs were further compared to the humans to show similarity.

Results: Our findings indicate that LLMs not only match but, in several aspects, surpass even highly educated humans in extracting information, including both demographic and context details, from unstructured texts.

Conclusions: This study not only validates the use of LLMs for analyzing digital healthcare communications but also opens new avenues for understanding online behaviors and interactions, signaling a shift towards more sophisticated methodologies in digital research and practice.

(JMIR Preprints 17/03/2025:74094)
DOI: https://doi.org/10.2196/preprints.74094

Original Manuscript

Assessing Large Language Models in Building a Structured Dataset from Reddit Data: A Methodological Study

Chase Westhoff, John Westhoff*, Ethan Low, Carl Hanson, E. Shannon Tass, Quinn Snell

Brigham Young University

*University of Nevada Reno

Abstract

Background: In an era marked by the blooming reliance on digital platforms for healthcare consultation, the subreddit r/AskDocs has emerged as a pivotal forum. However, the vast, unstructured nature of forum data presents a formidable challenge; the extraction and meaningful analysis of such data require advanced tools that can navigate the complexities of language and context inherent in user-generated content.

Objective: Our objective was to evaluate employing Large Language Models (LLMs) to systematically transform the rich, unstructured textual data from AskDocs into a structured dataset, an approach that aligns more closely with human cognitive processes compared to traditional data extraction methods.

Methods: We developed a dataset of Reddit posts from r/AskDocs by extracting key information via human annotators. Then using specially engineered prompts we used state-of-the-art Large Language Models (LLMs) to extract data from posts and compared the results. The variation in the LLMs were further compared to the humans to show similarity.

Results: Our findings indicate that LLMs not only match but, in several aspects, surpass even highly educated humans in extracting information, including both demographic and context details, from unstructured texts.

Conclusions: This study not only validates the use of LLMs for analyzing digital healthcare communications but also opens new avenues for understanding online behaviors and interactions, signaling a shift towards more sophisticated methodologies in digital research and practice.

Keywords: Large Language Models; unstructured text; data extraction

Introduction

The advancement of digital healthcare, especially highlighted during the COVID-19 pandemic, has significantly increased the reliance on online platforms for medical consultation and advice, profoundly changing how individuals seek medical advice over the last two decades [16,29]. With a growing focus on platforms like Reddit for "Ask the Doctor" services, Reddit's r/AskDocs subreddit has become a vital forum for such interactions, growing to over 550k subscribers by January 2024 [20]. This growth trend, illustrated by a notable surge in user engagement starting in 2018, signifies the platform's evolving role as a trusted source for medical advice online. Recent studies on social media platforms like Reddit have highlighted active user engagement in health-related discussions, such as medication abortion [24] and dermatology [7]. The r/AskDocs subreddit has been a focal point for analyzing user demographics and health topic trends, showing a significant increase in user posts over time

[20] (see Figure 1). This trend towards asynchronous healthcare, where individuals engage with healthcare professionals and peers over digital platforms, underscores a shift in how medical advice is sought and dispensed in the modern era.

The potential to harness insights from these forums is immense, offering a unique window into patient concerns, misconceptions, and the public's health-seeking behaviors. However, the vast, unstructured nature of forum data presents a formidable challenge; the extraction and meaningful analysis of such data require advanced tools that can navigate the complexities of language and context inherent in user-generated content.

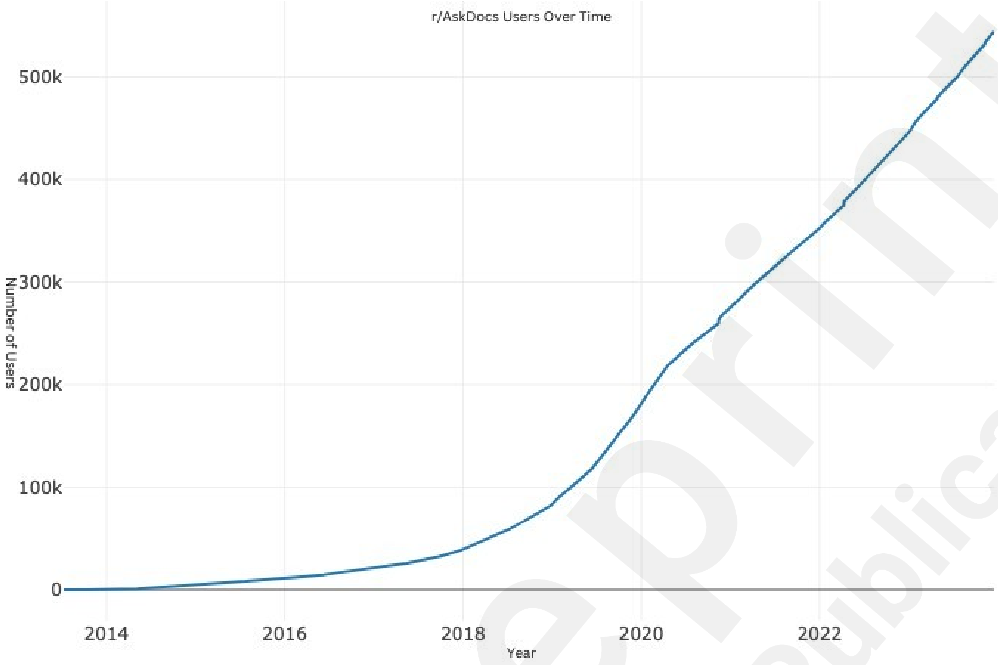


Figure 1: Growth of the subreddit r/AskDocs over time. [2]

Traditional Methods of Information Extraction from Text

Regular expressions (often shortened to regex), a staple in text processing, offer a rule-based approach to identifying specific patterns within text. For example, regex can be used to locate all instances of email addresses or phone numbers within a database by defining the patterns that match email addresses and phone numbers. In the realm of data extraction from online health forums such as AskDocs, the intricacy of human language and the unstructured nature of user submissions present significant challenges. Users frequently provide a wealth of information, albeit in varied formats that defy simple pattern matching. The diversity in the presentation of this data complicates the task of developing regular expressions that can accurately and consistently extract the desired information. [19]

To illustrate the complexity of this task, consider the following examples that represent patterns seen in posts on AskDocs. These examples highlight the variability and nuanced nature of the information provided by users, underscoring the difficulties in crafting Regex patterns capable of effectively parsing and categorizing this data.

Table 1: Examples of AskDocs User Submissions and Their Complexity

Text	Explanation
"I'm worried about a white lump on my elbow. Age: 31; Race: Japanese; Sex: Male"	The text clearly states their age and race, but uses "white" first in a medical context, not as a race.
"Mid-30s F here, experiencing severe headaches. Also, I'm 5 foot 6" and around 60-ish kg."	Use of approximate age ("Mid-30s") and nonstandard expressions for measurements ("5 foot 6"" and "60-ish").
"I'm a minor, dealing with severe migraines especially during my period. 150cm, 60kg"	The text implies the user's sex through the mention of a menstrual cycle but does not explicitly state it.

These examples underscore the inherent challenges in using regular expressions for data extraction from AskDocs. The variability in how users report their demographic information, combined with the use of language in medical contexts, necessitates a highly sophisticated system using a variety of regex patterns. Building such a system is not only daunting, but entirely impractical.

Large Language Models

In contrast, large language models (LLMs) introduce a paradigm shift in data extraction. Large language models trained on a vast corpora of text present an understanding of language nuances, context, and the implicit meanings embedded within text. This enables LLMs to interpret and categorize complex information without the need for explicitly defined rules, as is the case with regex. [13]

The integration of LLMs in analyzing online health forums represents a significant advancement in this field. Large language models trained on extensive datasets have proven effective in understanding and generating humanlike text. For instance, the GatorTron model, a large clinical transformer model, demonstrated remarkable performance in extracting and utilizing patient information from clinical narratives [30]. Furthermore, the effectiveness of LLMs in preserving privacy while extracting information highlights their growing importance in sensitive domains like healthcare [23].

In addition to their general capabilities, LLMs have shown particular proficiency in tasks like information extraction, categorizing text data, and identifying sentiments in complex and unstructured data settings [3]. Another study showed how fine-tuning LLMs like GPT3 can accurately extract complex scientific knowledge [11]. This makes them highly suitable for extracting nuanced information from health-related discussions online.

In summary, the existing research lays a comprehensive foundation for understanding online health-seeking behavior, with LLMs playing a crucial role in advancing the understanding and analysis capabilities in digital health communication.

Research Question and Aim

Amidst this backdrop, this study provides an evaluation on the methodology of leveraging LLMs to transform the unstructured, information-rich text data of sources such as the AskDocs subreddit into a structured dataset. Unlike traditional data extraction methods, which struggle with the variability and complexity of natural language, LLMs offer a context-aware, nuanced approach that more closely aligns with human cognitive processes. The following research question seeks to evaluate the methodology's effectiveness and explore its broader applications:

- RQ: How do the accuracy and agreement of LLMs in labeling online health communication compare to human annotators in extracting and categorizing complex information from AskDocs?

This study focuses on the methodology and its validation, aiming not only to demonstrate the feasibility of using LLMs for analyzing online health communication but also to examine the broader implications of this approach for digital research and practice. To show that our method produces usable data, we conducted a cursory analysis of the produced data as an example usage, illustrating its potential for future analysis. This approach seeks to uncover new avenues for understanding online behaviors and interactions.

While this research focuses on extracting structured datasets from health forum data, the applicability of LLMs extends far beyond this realm. Their versatility and advanced understanding of natural language make them suitable for various fields requiring data extraction and analysis. These fields could include, but are not limited to, legal document analysis, financial report summarization, and sentiment analysis in social media [10,15,25]. LLMs offer a powerful tool for transforming unstructured text into actionable insights. This broad applicability underscores the transformative potential of LLMs across multiple sectors, promising to revolutionize data analysis and knowledge extraction in an array of disciplines.

Methods

The process for data collection, human labeling, and labeling from different LLMs and the process for comparing label similarity between human annotators and LLMs is outlined below and illustrated in Figure 2.

Data Collection

The AskDocs subreddit is part of Reddit, which hosts over 13 billion posts across more than 100,000 subreddits, engaging more than 50 million daily users [1]. Due to limitations in Reddit's official API, PRAW, for comprehensive historical data retrieval, we leveraged the Pushshift API [5], a third-party archive of Reddit's post metadata. Despite known gaps in Pushshift's data coverage, it remains a valuable tool for accessing large volumes of Reddit data, including the AskDocs subreddit.

Data were extracted from AskDocs spanning from inception in July 2013 to October 2022, comprising 1,016,229 posts and 2,122,081 comments. While acknowledging the challenges in obtaining a complete dataset, the available data was substantial for extracting insights and addressing the research objectives of this study.

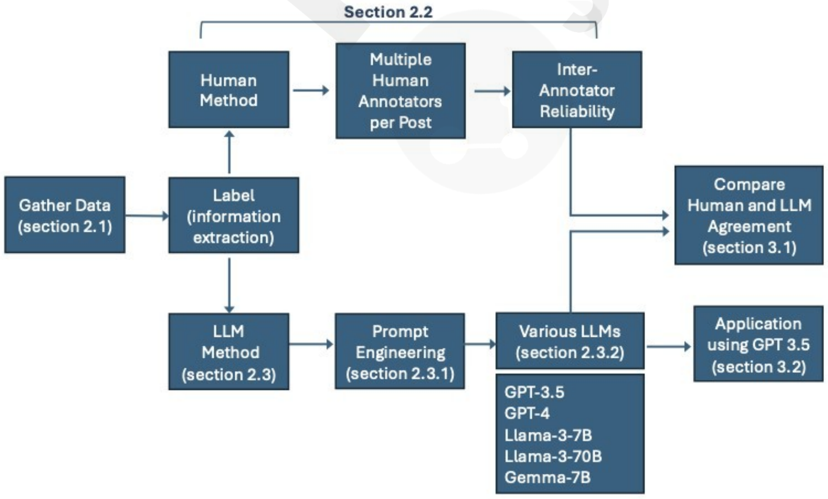

Section 2.2
Human Method
Multiple Human Annotators per Post
Inter-Annotator Reliability
Gather Data (section 2.1)
Label (information extraction)
Compare Human and LLM Agreement (section 3.1)
LLM Method (section 2.3)
Prompt Engineering (section 2.3.1)
Various LLMs (section 2.3.2)
Application using GPT 3.5 (section 3.2)
GPT-3.5
GPT-4
Llama-3-7B
Llama-3-70B
Gemma-7B

Figure 2: Flowchart of methods used for this study.

A random sample of 2800 posts was used for comparing data extraction using human labeling and LLM labeling. A different sample of approximately 30,000 posts was used to demonstrate the LLM data retrieval methodology.

Human Labeling

Human labeling in extracting and categorizing information was essential for this study for multiple reasons. First, it provided a benchmark for evaluating the accuracy and reliability of LLMs against the ”gold standard” of human cognition and understanding. Second, this comparison shed light on the potential of LLMs to augment or even surpass human efforts in terms of efficiency, scalability, and consistency. Understanding where LLMs excel or fall short compared to human annotators allows for better utilization of capabilities, identification of areas for improvement, and refinement of methodologies to enhance their performance in real-world applications.

Collecting Human Labels

In collaboration with the University of Nevada, Reno Medical School, 27 medical students were engaged to undertake the task of data labeling. The strategy involved randomly selecting 3,600 posts, organized into nine batches of 400 posts each, to provide a comprehensive snapshot of the prevalent dialogues on the platform. Each batch was reviewed by three medical students to ensure thorough examination of each post from multiple perspectives. The students were tasked with categorizing the posts according to the extraction criteria fields outlined in Table 2.

The execution of this plan encountered practical challenges. Despite our initial aspirations, only two of the nine batches saw the completion of labeling by all three assigned students. Five batches had the contribution of two students, and unfortunately, two batches were only labeled by a single student. Given a commitment to ensuring that each post was reviewed by at least two individuals to guarantee robustness and reliability of the human-labeled dataset, the decision was made to exclude the batches that were only reviewed by one student. As a result, the final dataset comprised 2,800 posts, each labeled by at least two medical students, with ties broken by a lead researcher. This dataset then served as the benchmark for a comprehensive comparison with the data extraction by various LLMs.

Table 2: Information fields and possible responses for human and LLM data extraction from the Reddit posts.

Field	Options
Biological Sex	M, F, Unknown, NA
Gender Identity	M, F, Other, NA
Age	A numerical value, Unknown, NA
Height	Height in formats like 6’0” or 170cm, Unknown, NA
Height Units	feet/in, cm, m, NA
Weight	A numerical value, Unknown, NA
Weight Units	lbs, kg, NA
Race	White, Asian, Black, Hispanic, Other, Unknown
Diagnosis Based Medical Inquiry	True, False, NA
Symptom Based Medical Inquiry	True, False, NA
Treatment Based Medical Inquiry	True, False, NA
Proxy Relationship [1]	NA, Significant Other, Friend, Child, Other

Chronic Condition	True, False, NA
Healthcare Consultation Status	Pre-consultation, in consultation, Post-consultation, NA
Primary Focus Topic	Multiple possible topics

Table 3: Interpretation of Cohen’s kappa as proposed by [18].

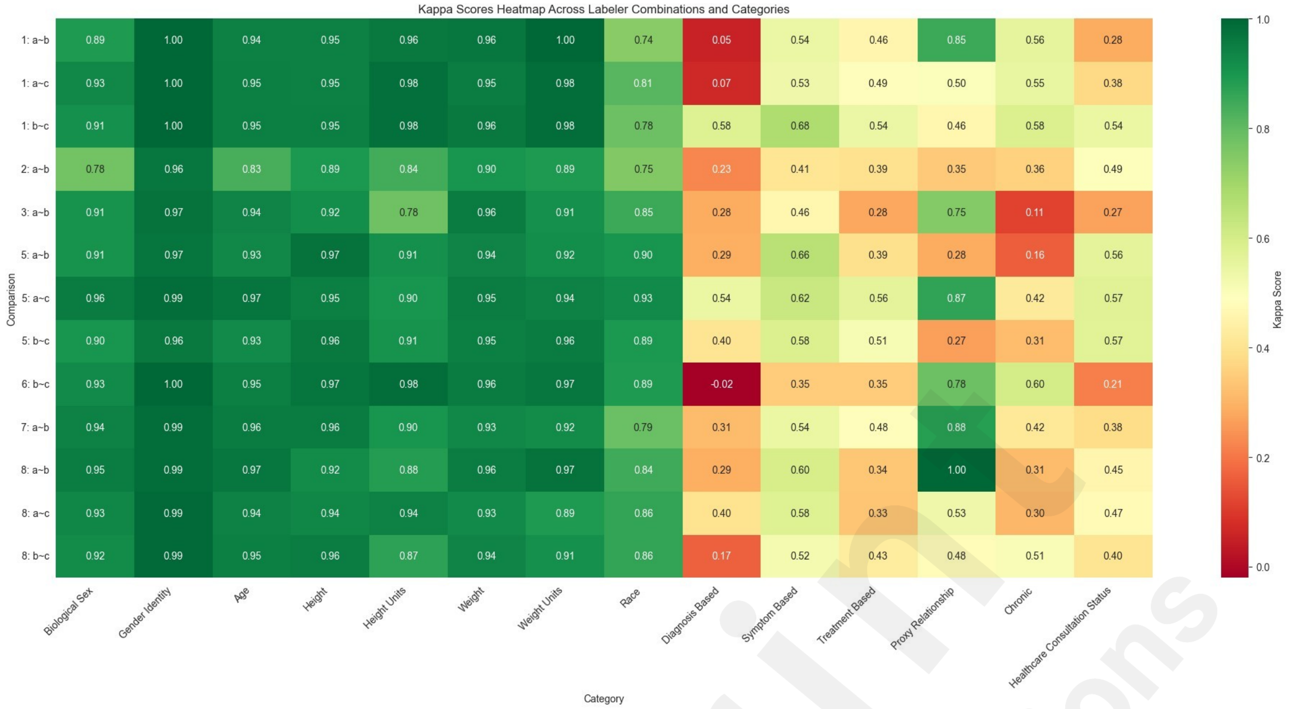
Kappa Score	Agreement Level
0-0.2	Almost none
0.21-0.39	Minimal
0.40-0.59	Weak
0.60-0.79	Moderate
0.80-0.90	Strong
Above 0.90	Almost Perfect

Human Inter-Annotator Agreement

Cohen’s kappa score, a statistical measure for evaluating the level of agreement between two or more annotators, was utilized to assess inter-annotator reliability. This measure, introduced by Jacob Cohen [9], accounts for the possibility of chance agreement in its calculation, making it a more robust indicator of inter-annotator reliability than simple percent agreement. The magnitude of the kappa score indicates the level of agreement, with a score of 1 corresponding to perfect agreement. There are different ways to categorize agreement level using the kappa score. One such categorization proposed by McHugh [18] is shown in Table 3.

In the realm of data labeling, particularly for qualitative data with subjective categories, the Cohen’s kappa score is a standard metric for measuring annotator consistency. We illustrate the inter-annotator agreement with a Cohen’s kappa score matrix, showcasing the degree of consensus among different pairs of annotators on various categories.

Reliability in the current study is rooted in the accuracy of the human-labeled dataset, which is used as a benchmark and gold standard for evaluating LLMs performance. The Cohen’s kappa scores derived from the human annotators sheds light on an incurable aspect of human-mediated data labeling: the inherent diversity in human judgment. While high levels of agreement in some categories (left side of Figure 3) validate the clarity of our guidelines, the variability in others (right side of Figure 3) reflects the natural divergence in human interpretation. This phenomenon serves as a crucial reminder that disagreements between our benchmark dataset and LLMs generation of labels do not inherently signify an error on the LLMs part; rather, they may simply be a reflection of the educated guesses that humans often make in the face of ambiguity.



Comparison	Biological Sex	Gender Identity	Age	Height	Height Units	Weight	Weight Units	Race	Diagnosis Based	Symptom Based	Treatment Based	Proxy Relationship	Chronic	Healthcare Consultation Status
1: a~b	0.89	1.00	0.94	0.95	0.96	0.96	1.00	0.74	0.05	0.54	0.46	0.85	0.56	0.28
1: a~c	0.93	1.00	0.95	0.95	0.98	0.95	0.98	0.81	0.07	0.53	0.49	0.50	0.55	0.38
1: b~c	0.91	1.00	0.95	0.95	0.98	0.96	0.98	0.78	0.58	0.68	0.54	0.46	0.58	0.54
2: a~b	0.78	0.96	0.83	0.89	0.84	0.90	0.89	0.75	0.23	0.41	0.39	0.35	0.36	0.49
3: a~b	0.91	0.97	0.94	0.92	0.78	0.96	0.91	0.85	0.28	0.46	0.28	0.75	0.11	0.27
5: a~b	0.91	0.97	0.93	0.97	0.91	0.94	0.92	0.90	0.29	0.66	0.39	0.28	0.16	0.56
5: a~c	0.96	0.99	0.97	0.95	0.90	0.95	0.94	0.93	0.54	0.62	0.56	0.87	0.42	0.57
5: b~c	0.90	0.96	0.93	0.96	0.91	0.95	0.96	0.89	0.40	0.58	0.51	0.27	0.31	0.57
6: b~c	0.93	1.00	0.95	0.97	0.98	0.96	0.97	0.89	-0.02	0.35	0.35	0.78	0.60	0.21
7: a~b	0.94	0.99	0.96	0.96	0.90	0.93	0.92	0.79	0.31	0.54	0.48	0.88	0.42	0.38
8: a~b	0.95	0.99	0.97	0.92	0.88	0.96	0.97	0.84	0.29	0.60	0.34	1.00	0.31	0.45
8: a~c	0.93	0.99	0.94	0.94	0.94	0.93	0.89	0.86	0.40	0.58	0.33	0.53	0.30	0.47
8: b~c	0.92	0.99	0.95	0.96	0.87	0.94	0.91	0.86	0.17	0.52	0.43	0.48	0.51	0.40

Figure 3: Cohen's kappa score matrix displaying the agreement between different pairs of human annotators across the different categories of extraction. A higher kappa score indicates a stronger agreement.

High Disagreement Fields

As shown in Figure 3, the target questions exhibited varying levels of difficulty even for human annotators. The topics categorized as Treatment-Based, Diagnosis-Based, and Symptom-Based all aim to determine the reason for an individual's post. Treatment-Based questions pertain to discussions about treatments, Diagnosis-Based questions relate to diagnoses, and Symptom-Based questions focus on symptoms. However, these categories can sometimes be ambiguous, as demonstrated by the following post:

> "Dear doctors, I currently have a very minor case of poison ivy than that of my cases in the past; my girlfriend although believes it is contagious and refuses to make any contact with me (e.g. Hold my hand). Is poison ivy contagious? I am a male, 18, and have had poison ivy for about 4 days now."

In this instance, one annotator labeled the post as Diagnosis-Based, likely due to the mention of a poison ivy case, while the other two annotators classified it as Symptom-based, focusing on the symptoms of poison ivy.

Other challenging questions include those related to Chronic Conditions, Proxy Relationships, and Healthcare Consultation Status. Chronic Conditions questions inquire whether the issue is ongoing, Proxy Relationship questions address the poster's relationship to the person with the issue, and Healthcare Consultation Status questions indicate whether the individual has not yet consulted a doctor (Pre-Consultation), is currently consulting (In Consultation), or has completed the consultation (Post-Consultation).
Below are examples of posts that led to disagreements among annotators:

- "F23 test results came back and suggest a possible cyst. Should I pursue treatment? Here are the results. I originally went to the doc for lower back and side pain" (Consultation Status)

- “Are these bug bites, and if so, shold I be concerned? Here are some pictures of the bumps in question. Thepictures are two weeks old, and I still have them.” (Chronic Condition)
- “Do children need antibiotics for a uti? Writing about a 7F. Has been needing to urinate per every few minutes.Luckily no pain when urinating and no blood either. Can this go away with cranberry juice and lots of water or are antibiotics needed?” (Proxy Relationship)

As an indicator of the LLM’s effectiveness, we use accuracy against the majority rule of our medical student human annotators. Given the low agreement among human annotators on these topics, low LLM agreement reflects the inherent ambiguity within the dataset.

Large Language Model Labeling

Prompt Engineering

Prompt engineering is a critical process in the application of LLMs. It involves the careful design of prompts or instructions that guide the model in understanding and performing the desired task. The significance of prompt engineering lies in its ability to leverage the model’s inherent capabilities by translating the task at hand into a format that the model can comprehend and execute effectively, increasing the probability of receiving a correct response in the desired format. In the current study, the techniques employed in engineering prompts for this task were JSON (JavaScript Object Notation) fields formatting and few-shot prompting.

JSON Fields Formatting

To comprehensively capture information shared by AskDocs users, a set of fields was defined within a JSON structure, a common format for representing data in a clear an accessible manner. Each field was designed to hold specific types of information, with permissible values outlined to ensure consistency and accuracy in the extracted data. The structured nature of JSON facilitated for the straightforward combination of these fields into a cohesive dataset, where each post was transformed into a structured object. The same fields and possible values used by the human annotators were employed (see Table 2).

Few-Shot Prompting

Few-shot prompting with LLMs is an approach designed to enhance the models’ ability to perform specific tasks. Few-shot prompting involves creating a prompt that includes several examples of the task at hand (often in a Q/A or Input/Output pair format), followed by the task that needs to be done. This technique effectively ”primes” the model by providing it with a few examples of how to complete a specific task, thereby improving its ability to understand and execute similar tasks with new data. LLMs can perform few-shot prompting without finetuning, and Brown et al [12] showed that they can perform numerous Natural Language Processing (NLP) tasks when provided a few examples in its prompt.

For instance, if the task involves extracting demographic information from unstructured health forum posts, a few-shot prompt might include examples like:

> Example 1
>
> Input: “I’m a 34-year-old male experiencing frequent headaches.”
>
> Output: Age: 34, Gender: Male, Concern: frequent headaches.

Example 2

Input: “Female, 29, noticing a rash that appeared last week.”

Output: Age: 29, Gender: Female, Concern: rash appeared last week.

After presenting a few such examples, the model is then given a new, unseen piece of text and asked to perform the same task. This method capitalizes on the LLM’s ability to discern patterns and apply the learned extraction process to new data, enabling more accurate identification and categorization of information. Few-shot prompting thus represents a powerful tool in the prompt engineering toolkit, significantly enhancing the LLM’s utility for data extraction [4]. This technique is crucial for several reasons:

- Flexibility: The inherent advantage of employing few-shot prompting lies in its remarkable flexibility, circumventing the need for extensive model fine-tuning. Traditionally, adapting a model to perform a new or specific task necessitates a substantial investment in data labeling and computational resources for training, often making the process cost-prohibitive and time-consuming. Few-shot prompting, however, leverages the pre-existing knowledge and versatility of LLMs, enabling them to understand and execute tasks with just a handful of examples.
- Consistency: Few-shot prompting helps in standardizing the output format, improving the odds that the LLM generates data in the defined JSON structure.
- Accuracy in Information Extraction: Previous work ?? has shown that few-shot prompting of LLMs has the potential to drastically increase accuracy across a multitude of tasks of varying complexity.
- Proper JSON Field Creation: Although recent advancements allow enforcing JSON formatting through both OpenAI APIs and locally hosted models, these methods do not guarantee the generation of JSON objects with the correct fields. Few-shot prompting addresses this limitation by explicitly illustrating how each field should be populated, encouraging the model to produce objects with the appropriate field ’keys’.

In the current study, the prompt structure employed followed this structure:

1. A brief introduction to the task, clarifying the goal of converting unstructured text into structured JSON format.
2. Detailed instructions on how to approach the analysis, specifying the information that needs to be extracted and how it should be categorized into the JSON fields.
3. Examples to illustrate the labeling process, serving as templates for the LLM to follow. These examples demonstrate how to fill out each JSON field based on the content of the posts, ensuring clarity and precision in the output.

In our study, we experimented with both two-shot and seven-shot prompting techniques. Zhao et al. [27] demonstrated that while increasing the number of examples can improve the accuracy of results, the gains tend to diminish as more examples are added. Additionally, large language models (LLMs) are susceptible to majority label bias, where the output is biased towards labels that are more frequent in the prompt. To mitigate this bias, we carefully selected examples that included a diverse range of labels, minimizing repetition wherever possible.

By adopting few-shot prompting, the aim was to leverage the LLMs’ capabilities for consistent and accurate information extraction without expensive fine-tuning.

LLM Models

Four main types of models were used in the current study: open source models using Llama 3 from Meta [28] and Gemma from Google [26], along with proprietary models using GPT-3.5 [6] and GPT-4 [22] from OpenAI. The advantage of open source models is that they are free to use and the user has more control of how the data is used and stored. Proprietary models, on the other hand, are perceived to be more accurate.

Llama3 comes in different sizes with 8 billion parameters (7B), and 70 billion parameters (70B). Gemma is a 7 billion parameter (7B) model. Models with more parameters are usually more accurate, but they also require vastly more computing and storage resources to use. In comparison, GPT-3.5 has 175 billion parameters, and while there is no official disclosure on its size it can be assumed to be much larger.

The analysis used few-shot prompts containing two examples for all models except for GPT-3.5, which was run using both two examples and seven examples. Two examples were used for GPT-4 due to cost prohibitions associated with additional examples. The Llama3 and Gemma models were run using various sizes of context for comparison.

Results

LLMs Labeling Compared to Human Annotators

After running each model, the accuracy score was computed for the LLM results versus the benchmark human annotated data. Notably, Llama3 70B with seven few-shot prompt examples and GPT-4 with two few-shot prompt examples had the highest agreements with the benchmark data. Figure 4 shows the accuracy scores for all LLMs across each data field.

The top performing models were the largest, with GPT4 2-shot and Llama3 70B 7-shot having the highest overall accuracy. It appears that in general, GPT4 excelled in the simpler fields, such as Biological Sex, Gender Identity, and Age, while Llama3 70B performed marginally better in the high disagreement fields, with slightly higher scores in Diagnosis Based, Symptom Based, Treatment Based, and Chronic. However, this gap in performance could be due to the extra examples in the prompt.

One key finding was the performance of LLMs in more subjective areas, such as determining whether a condition was chronic or assessing the healthcare consultation status. Similar to human annotators, LLMs encountered challenges in these subjective categories, reflecting the inherent complexity and nuanced understanding required to make these determinations.

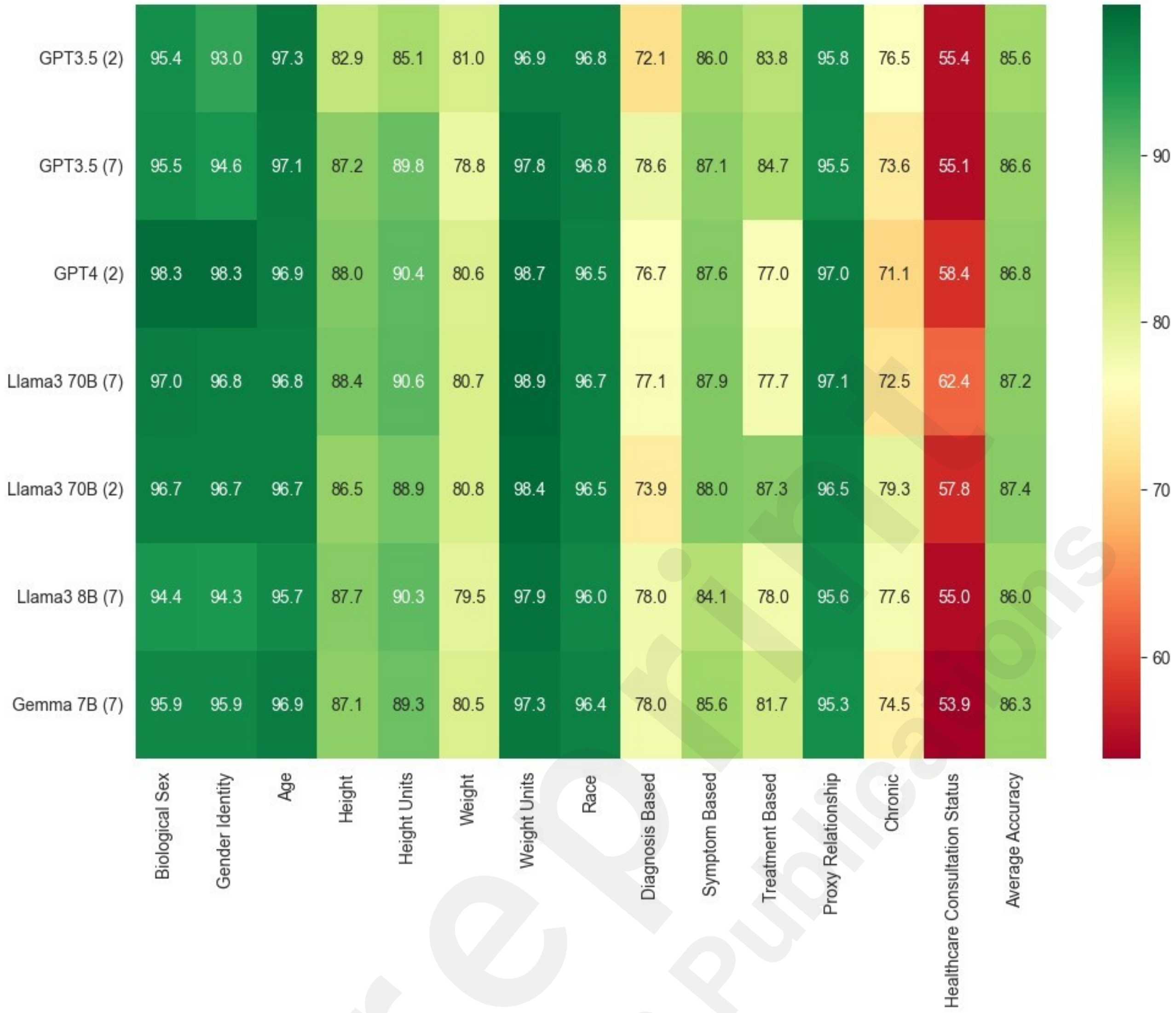


Figure 4: Detailed Performance Comparison of LLMs in Labeling AskDocs Posts Across Various Categories

Moreover, a side-by-side comparison of Cohen's kappa scores between the top two performing LLMs (GPT-4 and Llama3 70B) and a randomly selected pair of human annotators revealed a striking similarity in the pattern of disagreements. These results, shown in Figure 5, show that the differences between either LLM and Annotator A resembles the difference between Annotator A and B. This observation suggests that the discrepancies between LLM outputs and the benchmark dataset may mirror the natural variance found in human labeling efforts. This highlights the capabilities of LLMs to approximate human-like understanding and judgment in complex categorization tasks.

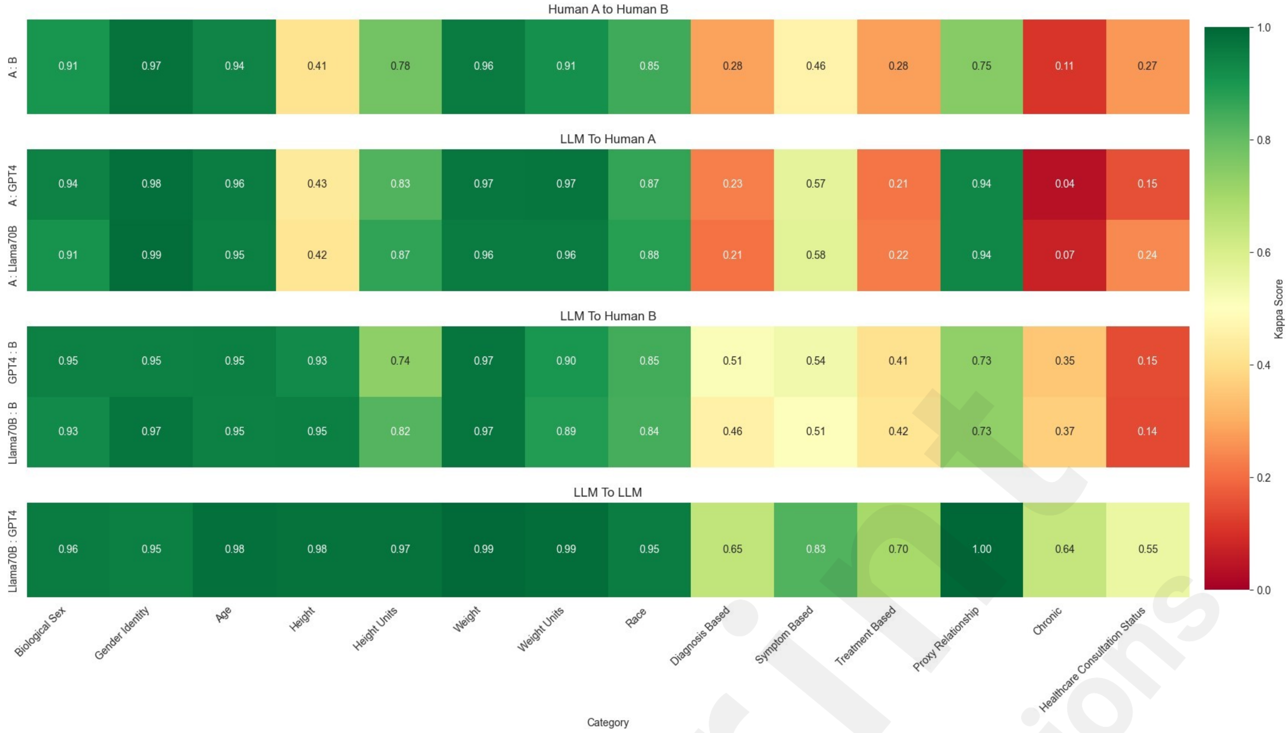


Figure 5: Cohen Kappa Scores: GPT3.5 vs. Benchmark Dataset and Between Two Human annotator

Demographic Insights and Health Discourse Trends

With the methodology established, it was applied to approximately 30,000 randomly sampled Reddit posts using GPT-3.5. This large-scale analysis offered unique insights into the health discourse on the AskDocs subreddit related to age distribution, the nature of inquiries, proxy relationship posts, and health topics discussed.

Age Distribution

Figure 6 illustrates the age distribution of users partaking in the AskDocs subreddit. The visible skew towards a younger demographic may reflect a generational trend in utilizing online platforms for health-related guidance. This observation aligns with broader usage patterns on Reddit, where 44% of users are aged between 18 and 29, and 31% are between 30 and 49 [17], suggesting that the demographic trends observed in AskDocs may indeed be representative of the general Reddit user base.

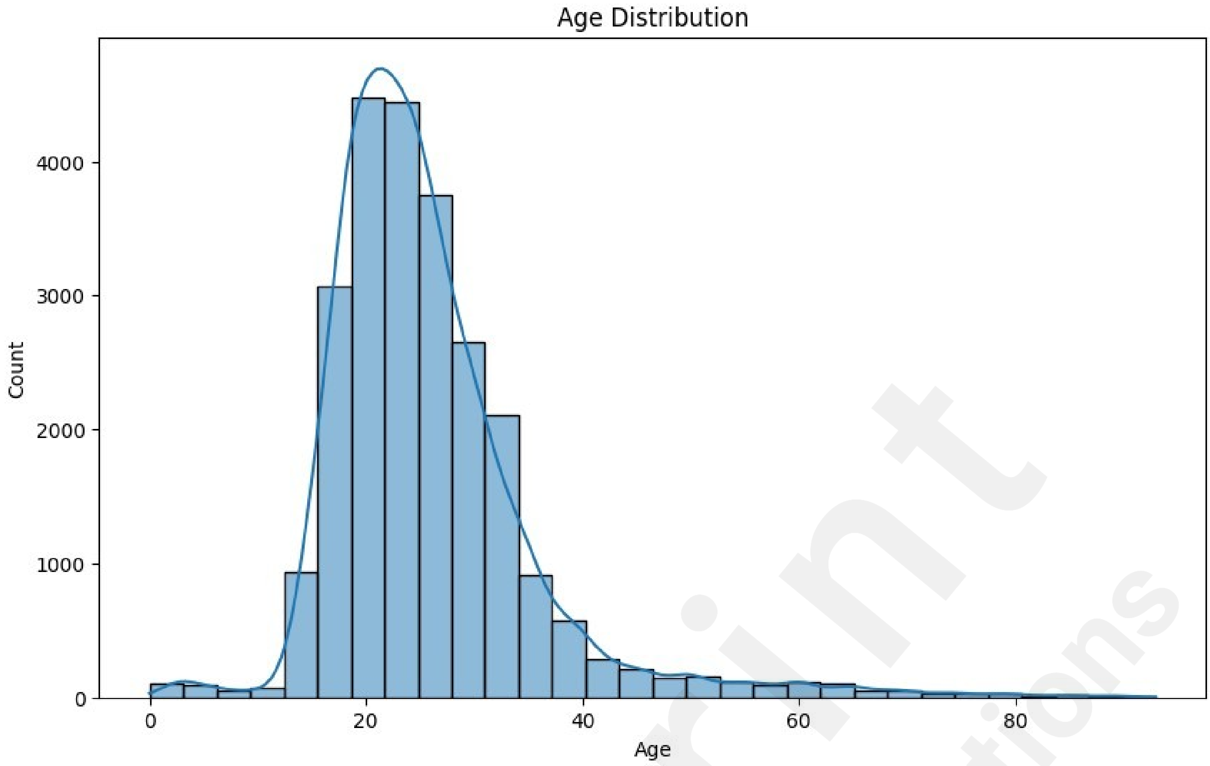


Figure 6: Age Distribution of AskDocs Users

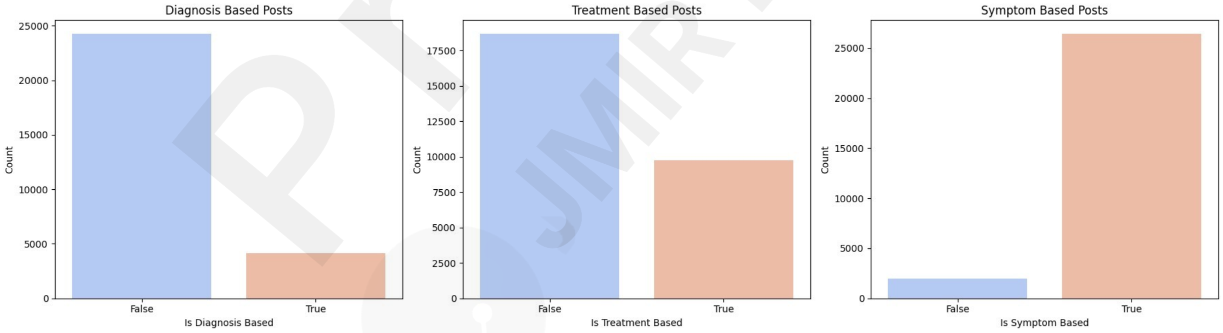


Figure 7: Categorization of Posts by Type of Medical Inquiry

Nature of Inquiry

In Figure 7, posts are segmented by the nature of the inquiry: Diagnostic-Based, Treatment-Based, or SymptomBased. The predominance of Symptom-Based queries suggests that users are often at an initial stage of seeking health information, which may involve symptoms checking before formal medical consultation.

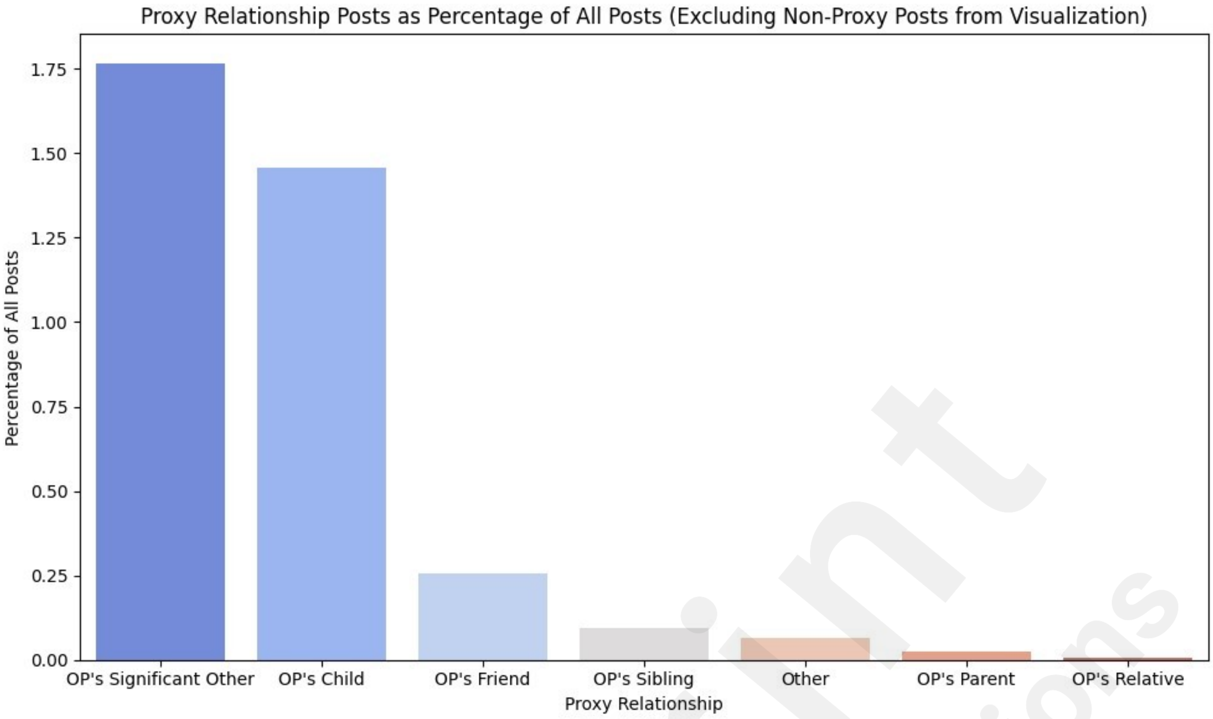


Figure 8: Proxy Relationship Posts as a Percentage of Total Posts

Proxy Relationship Posts

The vast majority (95%) of inquiries are made by individuals concerning their own health, emphasizing the personal use of AskDocs for health concerns. When queries are made on behalf of others, they are predominantly for significant others or children (see Figure 8).

Health Topics

Finally, Figure 9 details what percentage of posts pertained to each of the top 10 most frequently discussed health topics. Notably ’Anxiety’ and ’Respiratory Infections’ both saw marked increases in discussion volume from 2019 to 2020, likely influenced by the COVID-19 pandemic, reflecting public health trends and possibly exacerbated public anxieties [14].

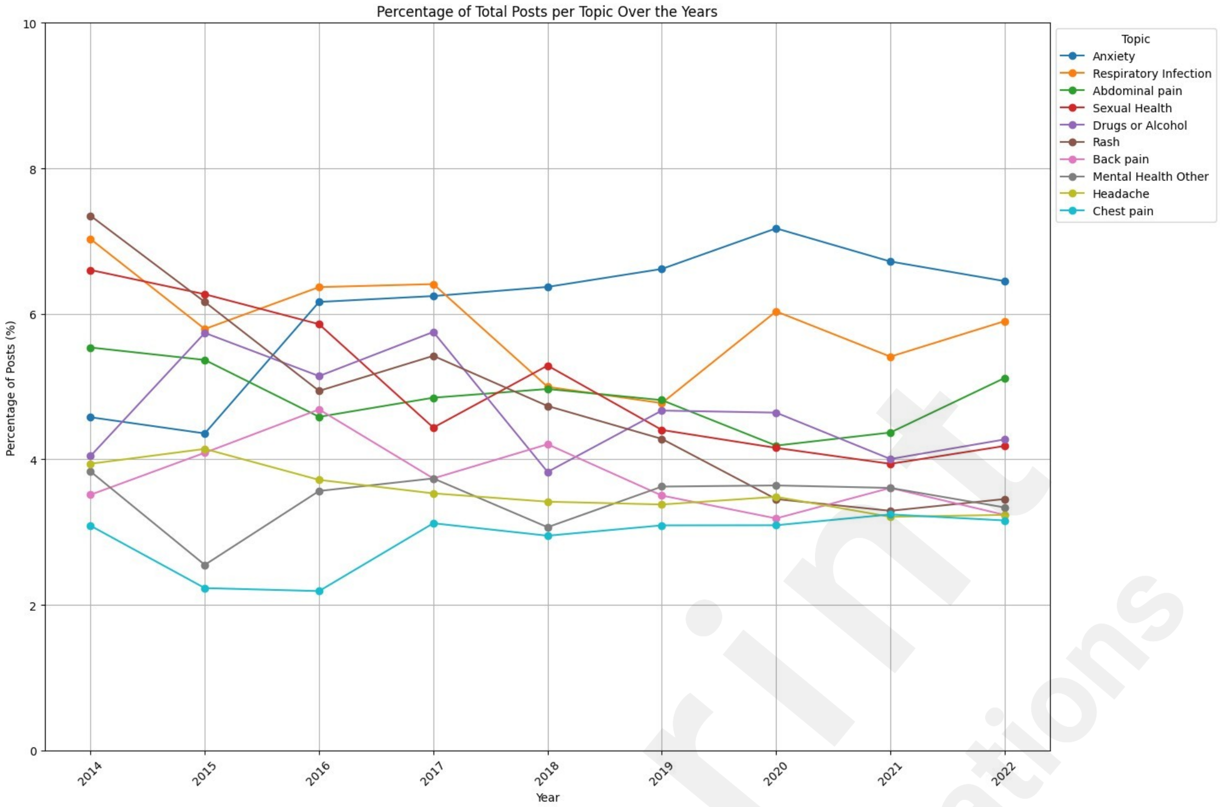


Figure 9: Most Common Health Topics Discussed on AskDocs as a Percentage of All Posts Over Time

Discussion

This study aimed to assess the application of LLMs in systematically converting the rich, unstructured textual data from the Reddit r/AskDocs subreddit into a structured dataset, a method that more closely mirrors human cognitive processes compared to conventional data extraction techniques. This comparative analysis shed light on the efficacy of labeling via LLMs relative to that of human annotators in the nuanced domain of an online health forum such as Reddit. The insights garnered point towards both the strengths and limitations of current AI technologies in domainspecific content understanding, paving the way for further research and development in the field of digital health communication.

Then, to demonstrate the potential for the LLM labeled data, a cursory analysis was performed that revealed patterns and trends within the r/AskDocs community. Insights like these have the potential to guide public health research, tailor online medical advice services, and support targeted health information dissemination, though further validation across more diverse datasets and additional forums is necessary for broader applicability and verification of these results.

The findings from this study shown in Figure 4 indicate that Llama3 70B with a seven few-shot prompt and GPT4 with a two few-shot prompt had the highest agreement with benchmark human-annotated data among the models run. Due to financial constraints we were unable to run GPT-4 with seven few-shot prompt. These results reflect the advanced capabilities of these models to understand and processing complex health-related information. Few-shot examples may enhance the performance of LLMs by improving their ability to recognize specific patterns, as they help the model interpret tasks more accurately [12] [8].

It is important to note that Llama3 70B is an open-source model, which allows it to be downloaded and run locally without incurring additional costs. This feature becomes particularly significant when considering information privacy or scenarios where data sensitivity prohibits the use of online servers. Furthermore, the performance of the open-source Llama3 70B is comparable to its proprietary counterpart, thereby enabling the application of these techniques in research contexts where such resources might otherwise be unavailable.

The pattern of disagreement between LLMs and human annotators exhibited notable similarities. Fields where human annotators had low Cohen's Kappa scores, as depicted in Figure 3, also posed challenges for LLMs. This suggests that the complexity of the target questions may contribute to the reduced LLM agreement due to human disagreement in the annotated dataset. For instance, fields with high disagreement among human annotators, as evidenced by low Cohen's Kappa scores, also showed low agreement between human annotators and LLMs, as shown by decreased accuracy. Conversely, fields where humans achieved high Cohen's Kappa scores, such as Biological Sex, also demonstrated high accuracy from LLMs.

This notion is further supported by the results illustrated in Figure 5. The comparison of Cohen's Kappa scores between our top-performing LLMs and a randomly selected pair of human annotators revealed that the pattern of disagreement between LLMs and human annotators mirrored the disagreement among the annotators themselves. This suggests that the consistency of LLMs is comparable to that of human annotators. It further indicates that LLMs can approximate human judgment, although perfect coding remains unlikely for subjective categories for both LLMs and humans. Therefore, when human annotators disagree with similar consistency as an LLM does with them, it may be reasonable to consider the LLM annotations with the same weight as those made by human annotators.

Cost and Time Efficiency of LLMs

The utilization of LLMs for data extraction, especially within the domain of healthcare and online forums such as r/AskDocs, brings forth numerous advantages that surpass traditional methods and human efforts in terms of efficiency, scalability, and applicability to a broader spectrum of fields. One of the most significant advantages of LLMs lies in their remarkable efficiency and speed in processing vast amounts of data. Unlike human annotators, who require considerable time to read, understand, and categorize information, LLMs can analyze and extract data from thousands of documents in a fraction of the time. This rapid processing capability is invaluable in settings where time-sensitive data extraction is critical, such as monitoring health forums for emergent public health concerns or extracting patient information from clinical notes in real-time.

LLMs exhibit a high degree of reliability and consistency in data extraction tasks. While human performance may vary due to factors such as fatigue, subjective interpretation, or inconsistencies in understanding, LLMs maintain a uniform standard throughout their operation. They follow the defined criteria and patterns with precision, reducing the variability and errors associated with human judgment. This consistency ensures that the extracted data is of uniform quality, facilitating more accurate analyses and decisions based on the information.

The scalability of LLMs significantly surpasses traditional data extraction methods and human labor. As the volume of data grows, especially with the ever-increasing reliance on digital health platforms, the need for scalable solutions becomes paramount. LLMs can be parallelized and deployed across multiple servers, allowing for simultaneous processing of data from various sources. This scalability enables researchers and organizations to handle data extraction tasks of any size, from analyzing a few documents to processing information across millions of online forum posts or clinical records.

Another significant advantage of using LLMs for data extraction is their inherent ability to maintain anonymity and preserve privacy. LLMs can be designed to process and extract relevant information without retaining any of the sensitive data, ensuring that the confidentiality of personal health information is maintained [31]. This aspect is particularly important in healthcare, where compliance with regulations such as HIPAA is essential. The ability of LLMs to ”forget” the specifics of the data they process allows for the extraction of valuable insights without risking the exposure of sensitive information.

It’s crucial to evaluate the economic and operational efficiency of using LLMs for large-scale data processing tasks. The costs and time required based on the current OpenAI API’s Tier 5 rate limits and token pricing [21] are shown in Table 4 and 5. Given the maximum context usage for efficient few-shot prompting, the total number of tokens per post, including inputs and outputs, is calculated as follows:

For each post, the input size is 16,000 tokens, and the output size is 1,000 tokens, resulting in a total of 17,000 tokens per post. For 30,000 posts, this totals 510,000,000 tokens.

Human labeling rate is 1 post per minute, taking 500 hours for 30,000 posts at a cost of $0.25 per post: - Total Cost: $7,500

Table 4: Cost comparison for labeling 30,000 posts using different methods.

Method	Cost Per Post	Total Cost
Human	$0.25	$7,500
GPT-3.5 Turbo	$0.0095	$285
GPT-4	$0.54	$16,200
LLama	Free	Free

Table 5: Operational time comparison for labeling 30,000 posts by different methods.

Method	Posts Per Minute	Total Time (Hours)
Human	1	500
GPT-3.5 Turbo	117	4.27
GPT-4	18	28.33
LLama3	19	26.31

The substantial cost and time efficiencies of GPT-3.5 Turbo compared to both human annotator and the more expensive GPT-4 model are notable. The ability to process large datasets quickly and economically with GPT-3.5 demonstrates its advantage for users needing high throughput and cost-effectiveness, whereas GPT4 offers advanced capabilities at a higher expense. As the LLM landscape progresses, the costs associated with these technologies are expected to fluctuate, potentially making high-capability models like GPT-4 more accessible.

Limitations and Future Research

Despite the demonstrated strengths of utilizing LLMs in information extraction, this research is not without limitations, which pave the way for future research opportunities. Large language model research is a rapidly advancing field, with new models and techniques regularly emerging. The methodologies employed in this study could potentially be refined by integrating state-of-the-art models and approaches that have been developed since the time of our research. Future studies should continuously incorporate the latest advancements to enhance the accuracy, efficiency, and reliability of data extraction processes.

Furthermore, our study faced financial limitations that restricted our ability to fully utilize GPT-4 with seven nshot samples. Although data trends indicated that the models performed better with an increased number of n-shot samples, we were unable to experiment with GPT-4 to its fullest potential. Consequently, the results might have been enhanced if we had the resources to conduct more extensive testing with additional n-shot samples.

A significant area for future research lies in applying these methodologies to analyze HIPAA-protected data. Currently, accessing such data involves complex legal processes to ensure privacy and compliance. By processing this data through an LLM, it may be possible to effectively extract the information without any humans viewing protected information, thereby facilitating analysis that was previously hindered by legal and ethical constraints. Research exploring the extent to which LLMs can maintain data anonymity while still providing valuable insights would be highly beneficial.

The introduction of new models with more extensive context lengths (allowing for longer prompts) provides an opportunity to include more examples in few-shot prompting, which may improve the model's understanding and execution of the data extraction task. Investigating whether the incorporation of more examples enhances the model's performance would provide valuable insights into the few-shot learning capabilities of LLMs. This research could involve experimental comparisons between models with varying context lengths to determine the optimal number of examples for accurate data extraction.

While our study targets the medical domain, particularly the AskDocs subreddit, the methodologies employed can and should be validated in a broader range of domains. Future research should extend beyond health forums to encompass a wide array of fields, creating large-scale datasets and employing LLMs for data extraction in each context. Comparing the performance of LLMs to expert human annotators and established automated methods across these varied domains is essential. This expanded benchmarking will not only solidify our understanding of LLMs' practical limitations but also verify their reliability and adaptability to diverse applications. Such cross-domain validation will underscore the versatility of LLMs and inform their refinement for specialized tasks.

Conclusion

While this study has laid the groundwork for the use of LLMs in extracting structured data from health forums, there remains significant potential for further exploration. Continuous advancements in LLM technology, combined with rigorous research into their applications and implications, will undoubtedly enhance our understanding and contribute to the evolution of digital healthcare research.

Conflicts of Interest – none declared

References

1 Press - reddit. https://www.redditinc.com/press. Accessed: 2023-01-21.

2 Askdocs subreddit statistics, 2024. Accessed: 2024-04-02.

3 Monica Agrawal et al. Large language models are few-shot clinical information extractors. 2022.

4 Toufique Ahmed and Premkumar Devanbu. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ASE '22, New York, NY, USA, 2023. Association for

Computing Machinery.

5 Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839, May 2020.

6 Tom B. Brown et al. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

7 T. Buntinx-Krieg, J. Caravaglio, R. Domozych, and R.P. Dellavalle. Dermatology on reddit: Elucidating trends in dermatologic communications on the world wide web. *Dermatology Online Journal*, 23(7):13030/qt9dr1f7x6, 2017.

8 Youngjin Chae and Thomas Davidson. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*, 2023.

9 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

10 Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. Llms to the moon? reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web Conference 2023*, WWW ’23 Companion, page 1014–1019, New York, NY, USA, 2023. Association for Computing Machinery.

11 Alex Dunn et al. Structured information extraction from complex scientific text with fine-tuned large language models. *ArXiv*, 2022.

12 Brown et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

13 Mohammed Hassanin and Nour Moustafa. A Comprehensive Overview of Large Language Models (LLMs) forCyber Defences: Opportunities and Directions. *arXiv e-prints*, page arXiv:2405.14487, May 2024.

14 R. Kindred and G. W. Bates. The influence of the covid-19 pandemic on social anxiety: A systematic review. *Int J Environ Res Public Health*, 20(3):2362, 2023.

15 Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF ’23, page 374–382, New York, NY, USA, 2023. Association for Computing Machinery.

16 Devin M. Mann, Ji Chen, Rumi Chunara, Paul A. Testa, and Oded Nov. Covid-19 transforms health care through telemedicine: Evidence from the field. *Journal of the American Medical Informatics Association*, 27(7):1132– 1135, July 2020.

17 MarketingCharts. Percentage of u.s. adults who use reddit as of september 2023, by age group [graph]. https://www.statista.com/statistics/261766/share-of-us-internet-userswho-use-reddit-by-age-group/, 2024. Accessed: 2024-04-12.

18 Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

19 Louis G. Michael, James Donohue, James C. Davis, Dongyoon Lee, and Francisco Servant. Regexes are hard: Decision-making, difficulties, and risks in programming regular expressions. In *2019 34th IEEE/ACM International Conference on Automated*

Software Engineering (ASE), pages 415–426, 2019.

20 A.L. Nobles, E.C. Leas, M. Dredze, and J.W. Ayers. Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 464–475, 2020.

21 OpenAI. Openai api documentation. OpenAI Platform Documentation, 2024. Accessed: 2024-04-10.

22 OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report, 2024.

23 Richard Plant et al. You are what you write: Preserving privacy in the era of large language models. *ArXiv*, 2022

24 Hana Reissner, Madeline R Ponzio, Lindsay Nagatani-Short, Alondra Hurtado, and Brian Nguyen. Medication abortion experiences before and during the covid-19 pandemic: A content analysis of online reddit posts [a19]. Obstetrics & Gynecology, 139(1):6S–6S, 2022.

25 Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM ’24, page 4882–4889, New York, NY, USA, 2024. Association for Computing Machinery.

26 Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Riviere, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research` and technology. *arXiv preprint arXiv:2403.08295*, 2024.

27 Shi Feng Dan Klein Tony Zhaozhi, Eric Wallace and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, 2021.

28 Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. Accessed: 2024-04-10.

29 D. Valdes, L. Alqazlan, R. Procter, et al. Global evidence on the rapid adoption of telemedicine in primary care during the first 2 years of the covid-19 pandemic: a scoping review protocol. *Systematic Reviews*, 11(124), 2022.

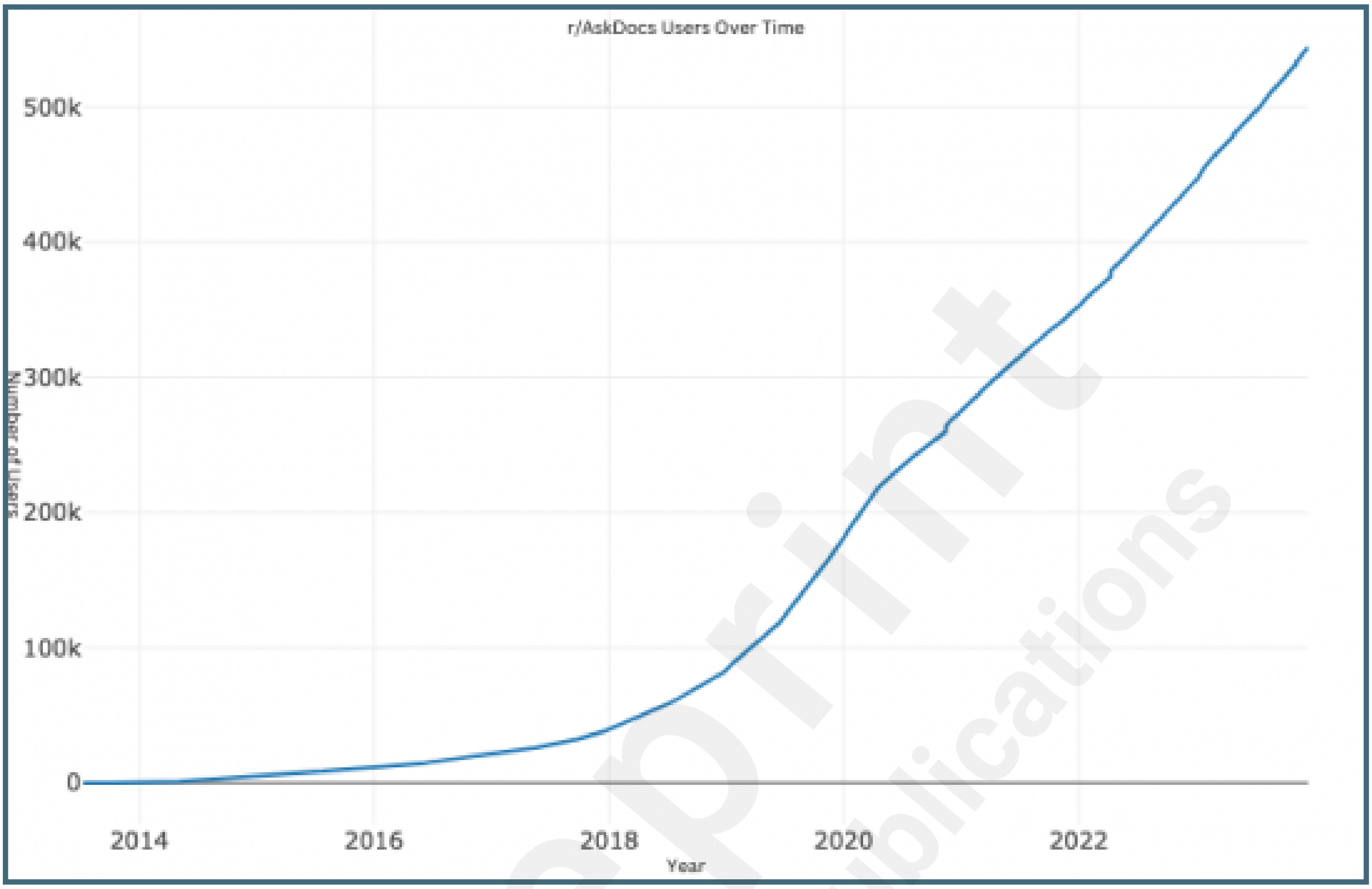
30 Xi Yang et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *ArXiv*, 2022.

31 Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.

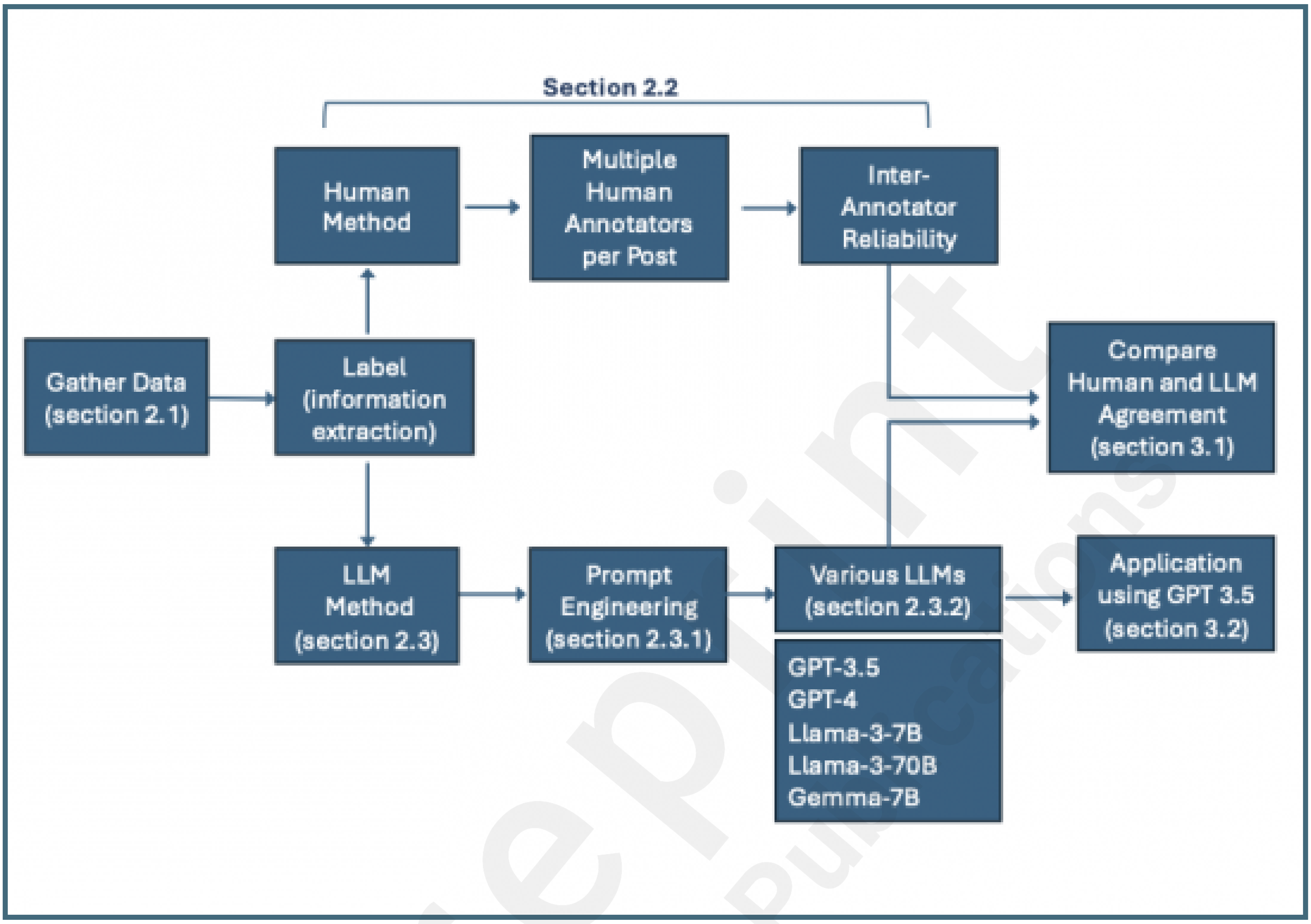
Supplementary Files

Figures

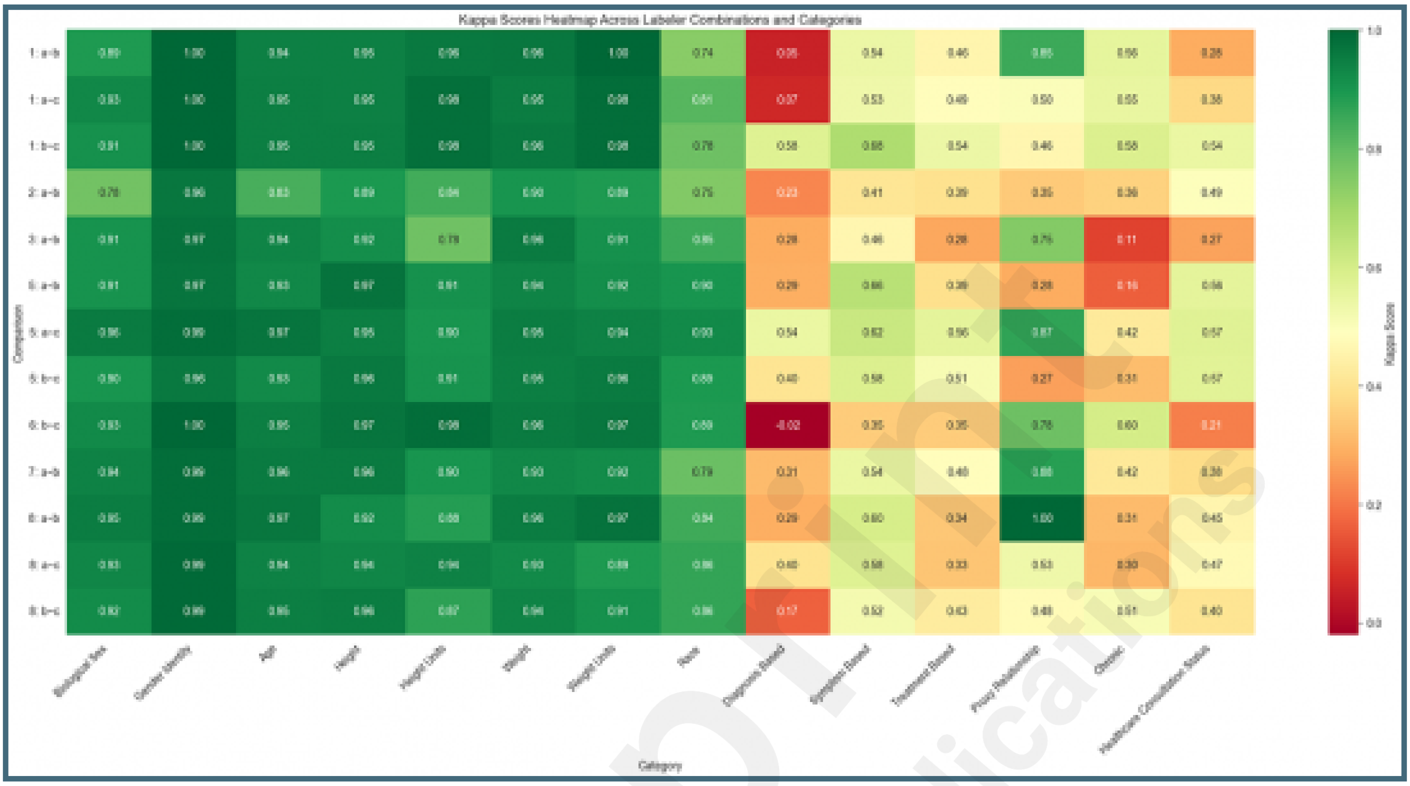
Growth of the subreddit r/AskDocs over time.



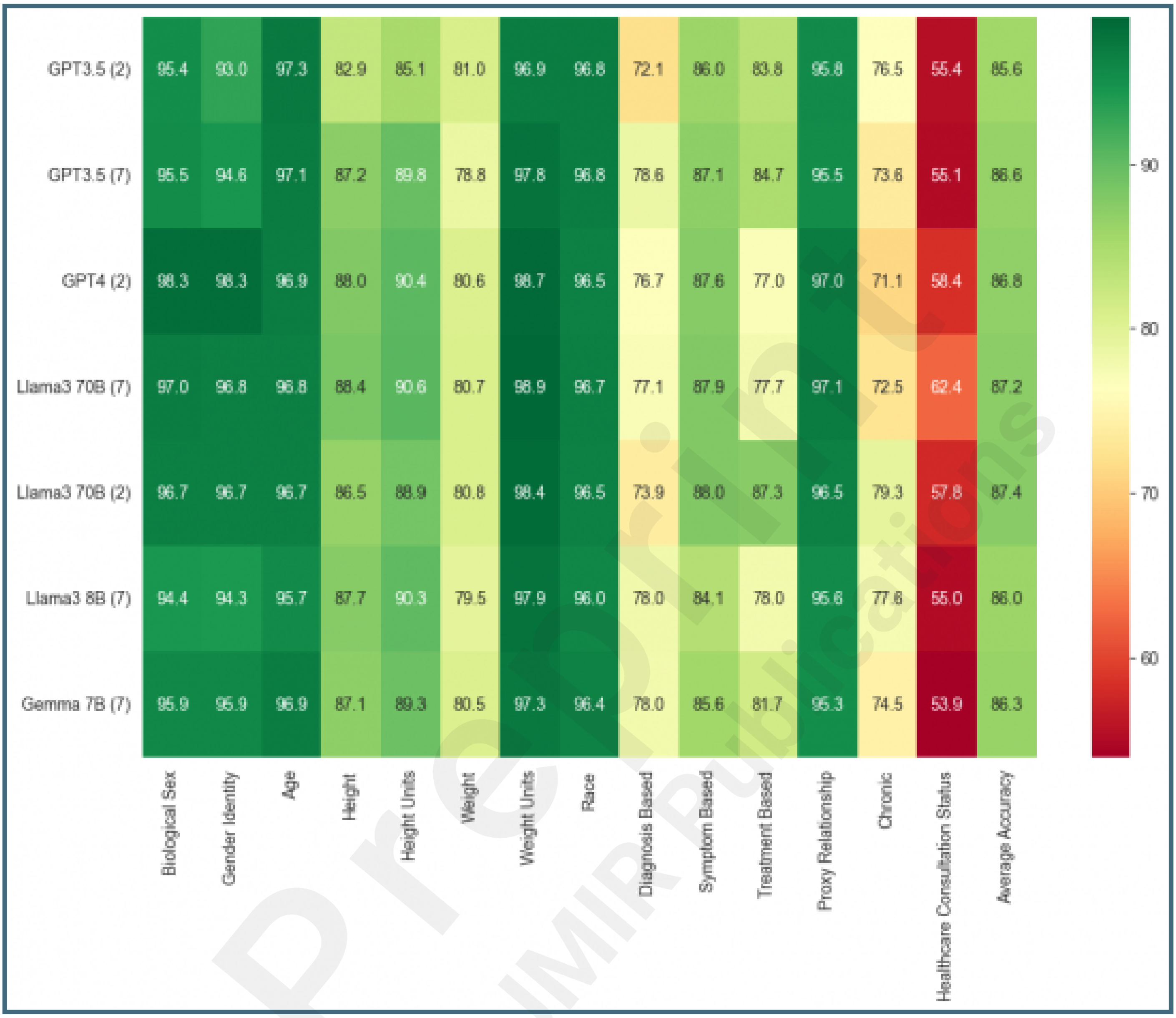
Flowchart of methods used for this study.


Section 2.2
Human Method
Multiple Human Annotators per Post
Inter-Annotator Reliability
Gather Data (section 2.1)
Label (information extraction)
Compare Human and LLM Agreement (section 3.1)
LLM Method (section 2.3)
Prompt Engineering (section 2.3.1)
Various LLMs (section 2.3.2)
Application using GPT 3.5 (section 3.2)
GPT-3.5
GPT-4
Llama-3-7B
Llama-3-70B
Gemma-7B

Cohen’s kappa score matrix displaying the agreement between different pairs of human annotators across the different categories of extraction. A higher kappa score indicates a stronger agreement.

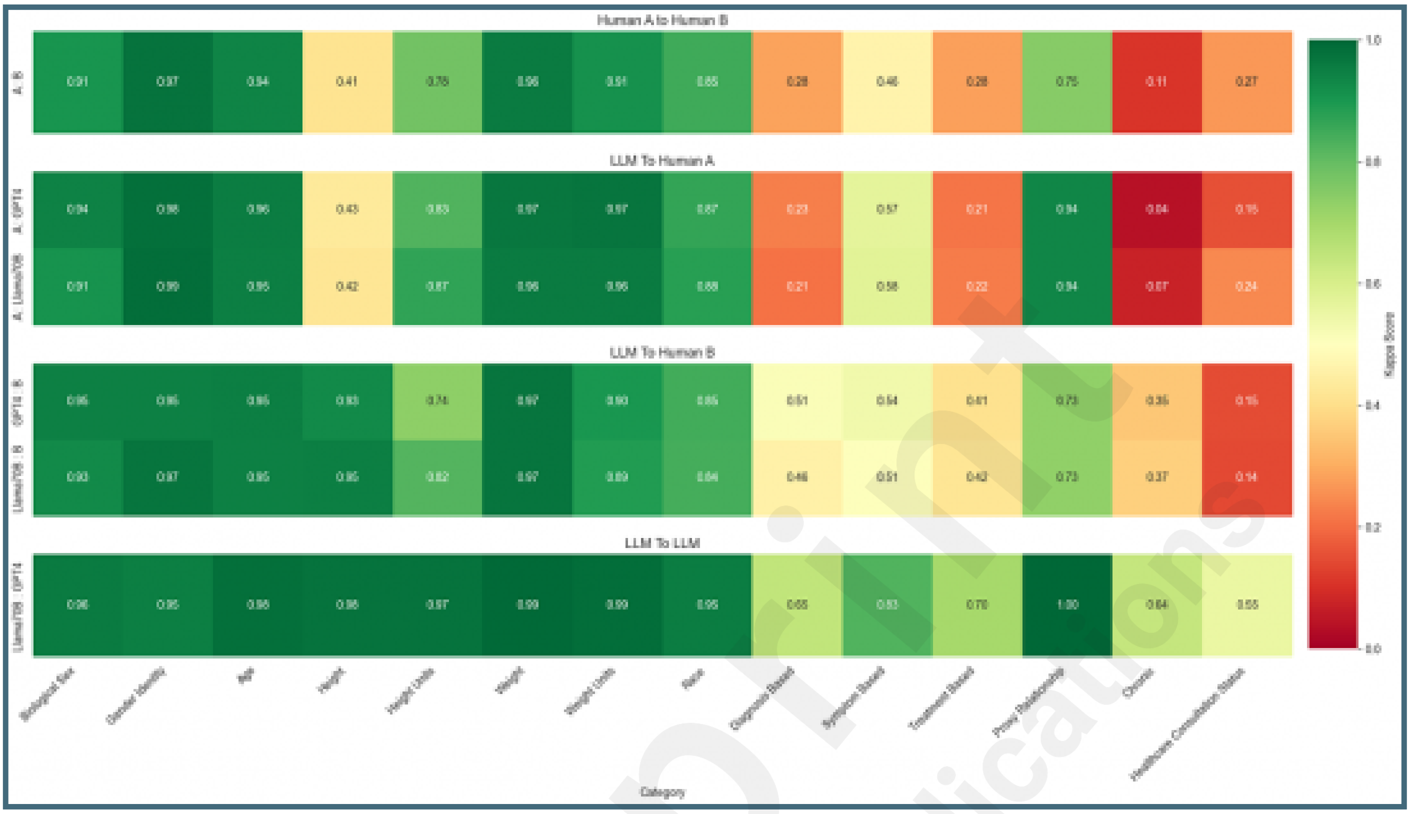


Detailed performance comparison of LLMs in labeling AskDocs posts across various categories.

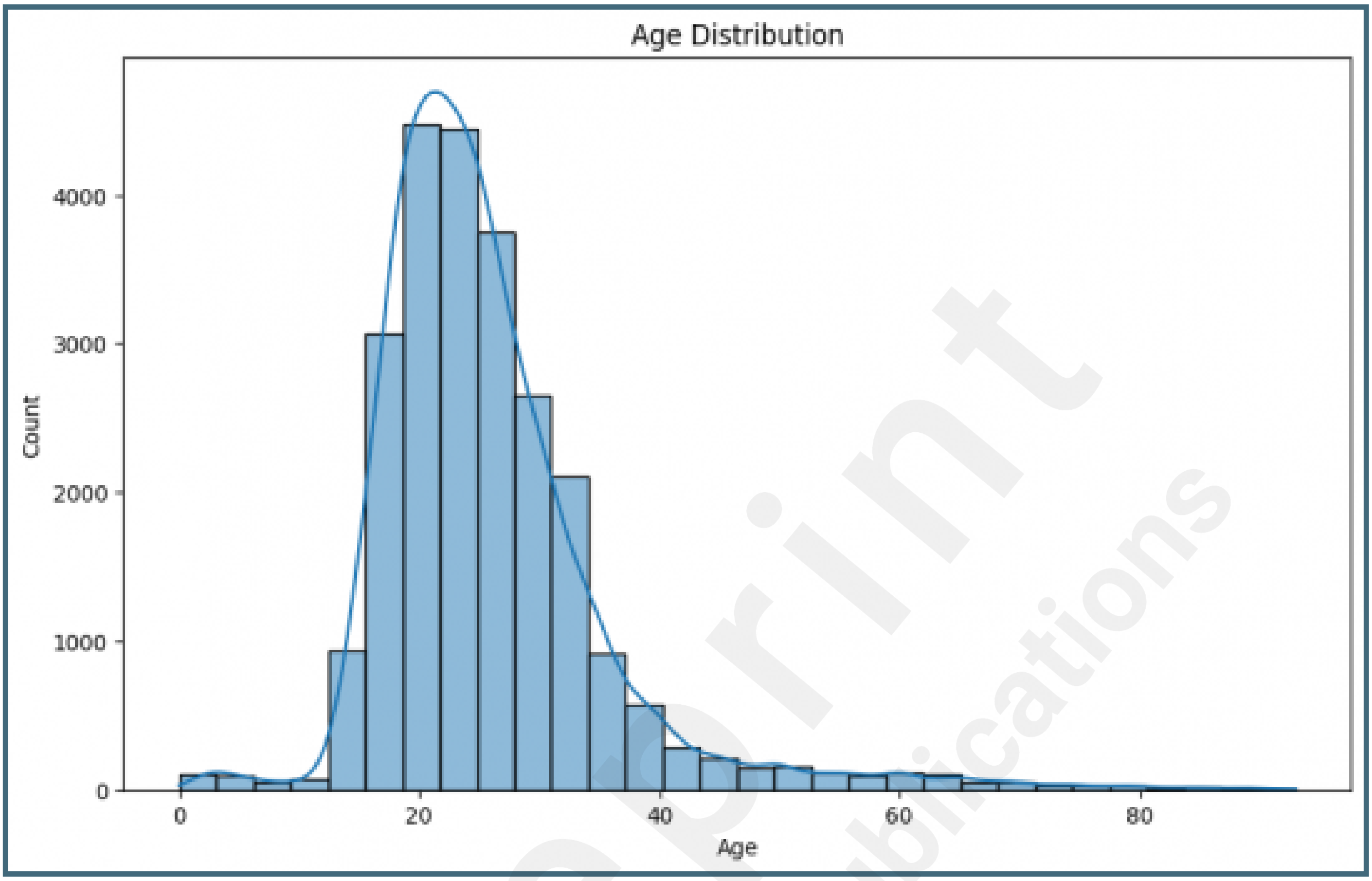


	Biological Sex	Gender Identity	Age	Height	Height Units	Weight	Weight Units	Race	Diagnosis Based	Symptom Based	Treatment Based	Proxy Relationship	Chronic	Healthcare Consultation Status	Average Accuracy
GPT3.5 (2)	95.4	93.0	97.3	82.9	85.1	81.0	96.9	96.8	72.1	86.0	83.8	95.8	76.5	55.4	85.6
GPT3.5 (7)	95.5	94.6	97.1	87.2	89.8	78.8	97.8	96.8	78.6	87.1	84.7	95.5	73.6	55.1	86.6
GPT4 (2)	98.3	98.3	96.9	88.0	90.4	80.6	98.7	96.5	76.7	87.6	77.0	97.0	71.1	58.4	86.8
Llama3 70B (7)	97.0	96.8	96.8	88.4	90.6	80.7	98.9	96.7	77.1	87.9	77.7	97.1	72.5	62.4	87.2
Llama3 70B (2)	96.7	96.7	96.7	86.5	88.9	80.8	98.4	96.5	73.9	88.0	87.3	96.5	79.3	57.8	87.4
Llama3 8B (7)	94.4	94.3	95.7	87.7	90.3	79.5	97.9	96.0	78.0	84.1	78.0	95.6	77.6	55.0	86.0
Gemma 7B (7)	95.9	95.9	96.9	87.1	89.3	80.5	97.3	96.4	78.0	85.6	81.7	95.3	74.5	53.9	86.3

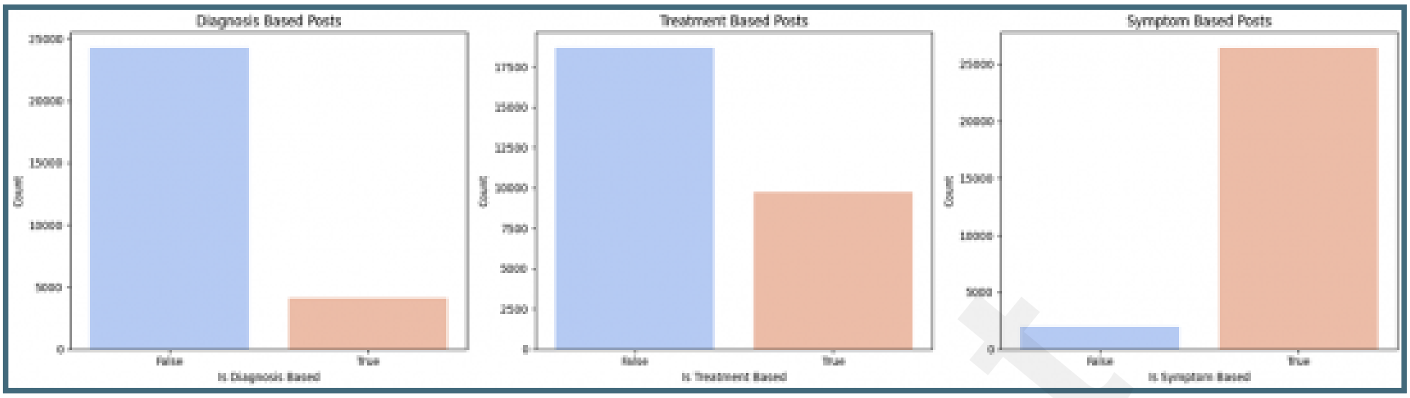
Cohen Kappa Scores: GPT3.5 vs. Benchmark Dataset and Between Two Human annotators.



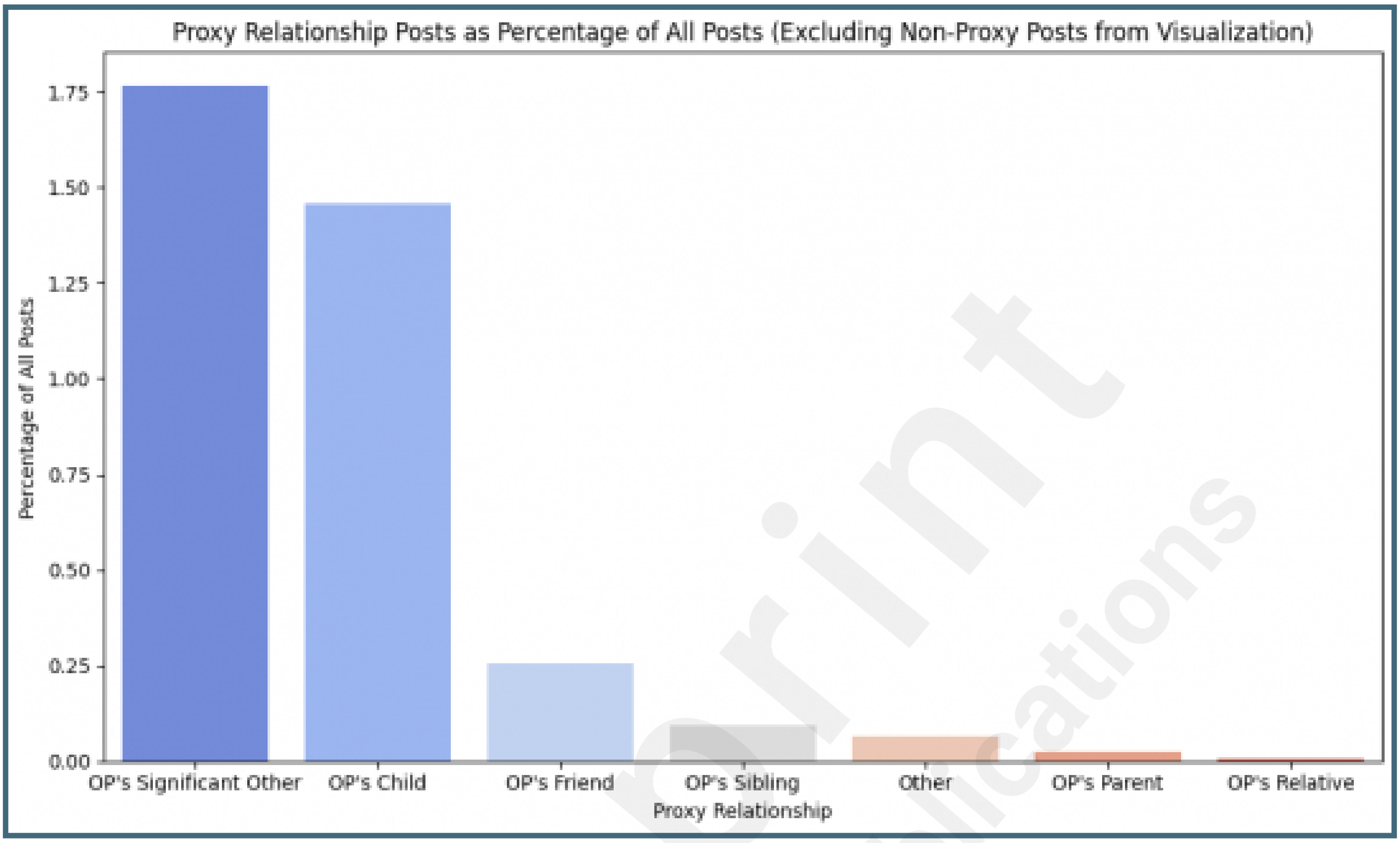
Age distribution of AskDocs users.



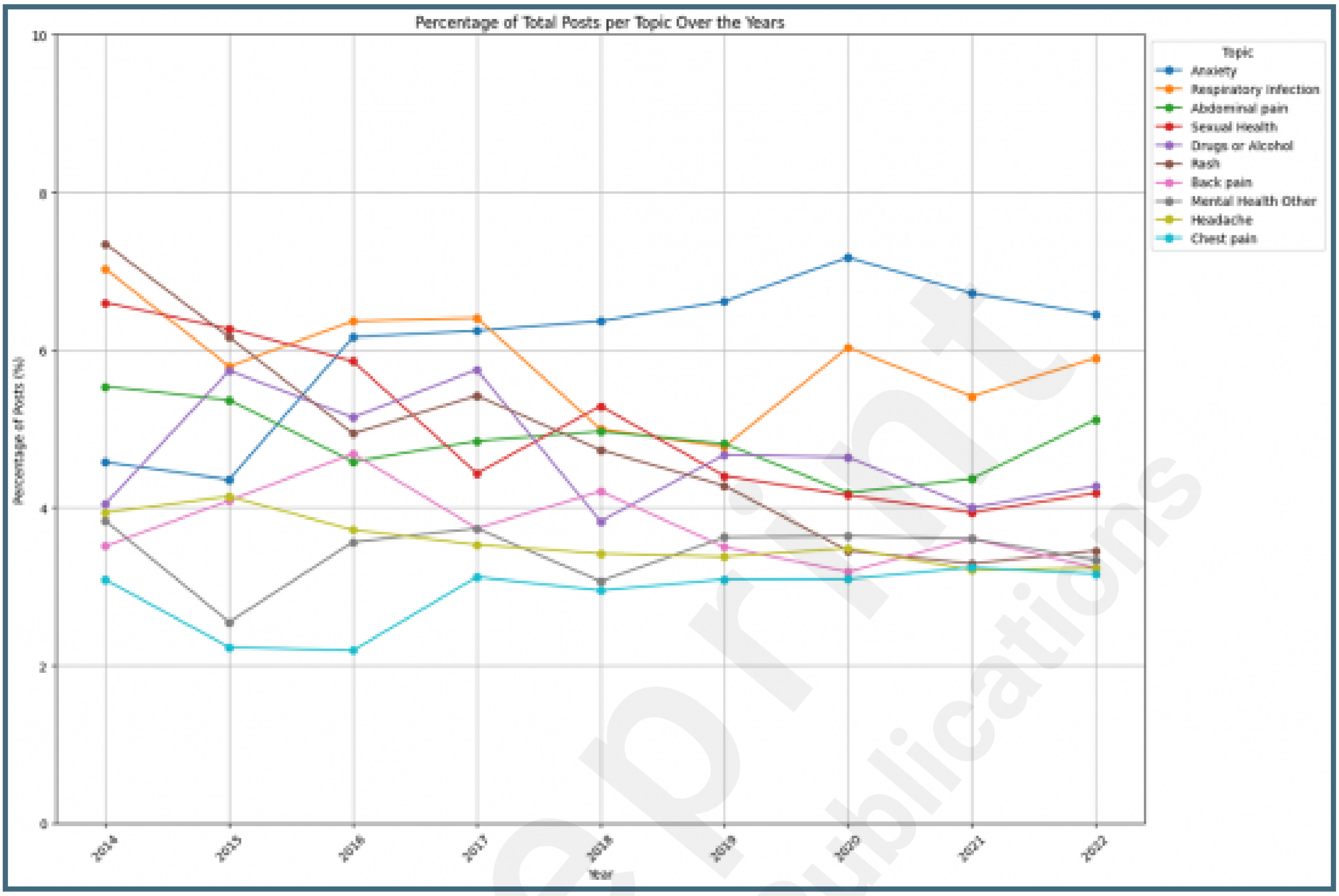
Categorization of posts by type of medical inquiry.



Proxy relationship posts as a percentage of total posts.



Most common health topics discussed on AskDocs as a percentage of all posts over time.


Percentage of Total Posts per Topic Over the Years
Percentage of Posts (%)
Year
Topic
Anxiety
Respiratory Infection
Abdominal pain
Sexual Health
Drugs or Alcohol
Rash
Back pain
Mental Health Other
Headache
Chest pain